Fundamenta Informaticae 34 (1999) 1–21
IOS Press

Contributions to the Theory of Rough Sets

V. Wiktor Marek and Mirosław Truszczyński*
Department of Computer Science
University of Kentucky
Lexington, KY 40506–0046
marek|mirek@cs.uky.edu

Abstract. We study properties of rough sets, that is, approximations to sets of records in a database or, more formally, to subsets of the universe of an information system. A *rough set* is a pair $\langle L, U \rangle$ such that L, U are definable in the information system and $L \subseteq U$. In the paper, we introduce a language, called the language of inclusion-exclusion, to describe incomplete specifications of (unknown) sets. We use rough sets in order to define a semantics for theories in the inclusion-exclusion language. We argue that our concept of a rough set is closely related to that introduced by Pawlak. We show that rough sets can be ordered by the *knowledge ordering* (denoted $\preceq_{kn}$). We prove that Pawlak's rough sets are characterized as $\preceq_{kn}$-greatest approximations. We show that for any consistent (that is, satisfiable) theory T in the language of inclusion-exclusion there exists a $\preceq_{kn}$-greatest rough set approximating all sets X that satisfy T. For some classes of theories in the language of inclusion-exclusion, we provide algorithmic ways to find this best approximation. We also state a number of miscellaneous results and discuss some open problems.

1. Introduction

In this paper we look at fundamental methodological issues underlying the concept of a rough set. Rough sets have been introduced by Pawlak [14] to serve as approximate descriptions of sets that are unknown, incompletely specified, or whose exact specification is complex. The approach pioneered by Pawlak allows us to reason about such sets given only their representations as rough

*This is an extended version of the first part of the presentation made by the authors at the RSCTC98, *Rough Sets and Current Trends in Computing*, an international meeting held in Warsaw, Poland, in June 1998. Address for correspondence: Department of Computer Science, University of Kentucky, Lexington, KY 40506–0046 sets. It found applications in databases, data mining, learning, approximate reasoning and many other areas of computer science. For more details on the theory and applications of rough sets we refer the reader to the monograph by Pawlak [15] and to several conference proceedings and collections of papers [23, 20, 13, 16, 17]. A good source of references is the Rough Sets web site `http://www.cs.uregina.ca/~roughset/`. Another useful reference is the the Bulletin of International Rough Set Society `http://www.cs.uregina.ca/~yyao/irss/bulletin.html`.

The study of rough sets is well motivated by practical applications. To illustrate the point, let us consider the following three scenarios.

1. The database language is inadequate to describe all subsets of some universe. This may happen when we want to reason about characteristics of objects represented by database records that become of interest after the database was designed and are not part of the language. For instance, in data mining we may be interested in consumer preferences with respect to a new group of products based on the past credit card data.
2. A set X of interest is unknown and we have only some information about it. We know about some sets that are disjoint with X and some that are included in X. We want to build good approximations to X and use them to reason about X. This situation occurs in many applications. For instance, in medicine a group of individuals at risk for a particular disease may be described in such terms. We want to be able to derive meaningful and correct, but not necessarily complete, information about people in this group.
3. The set X is known, and may even be definable in our database. Yet any description of X is so complex that it cannot be manipulated. In such case, we may have to use approximations of such a set that admit simple descriptions in order to be able to reason about it.

In this paper, we extend slightly the original definition of a rough set but change, quite dramatically, a perspective. Pawlak defined a rough set as an approximation to a *specific* set, say X. Pawlak's rough set corresponding to X is a pair $\langle \underline{X}, \overline{X} \rangle$ of two sets such that $\underline{X} \subseteq X \subseteq \overline{X}$. The sets $\underline{X}$ and $\overline{X}$ (defined with respect to a fixed information system — an issue to be made precise later in the paper) are called the lower and upper approximations to X. The emphasis on the set X, present in the original definition of a rough set, is what we strive here to free ourselves from. After all, in most (if not all) applications set X we want to reason about is unknown or is incompletely specified. Consequently, it may be that an approximation we use to reason about it is *not* its rough set (in the strict sense of Pawlak's definition).

In our proposal, the fundamental concept and the starting point is that of an approximation rather than that of a set to be approximated. This choice seems appropriate as approximations are *known* and can be reasoned about. An *approximation* is any pair of sets $\langle L, U \rangle$ such that $L \subseteq U$ (we will also require that these sets be definable in an information system). An approximation $\langle L, U \rangle$ serves as an approximation to *any* set X such that $L \subseteq X \subseteq U$. The main goal of our research is to study properties of approximations $\langle L, U \rangle$ and relate them to properties of sets X that they approximate.

It turns out that Pawlak's rough sets and our approximations are closely related. Clearly, all Pawlak's rough sets are approximations. More interestingly, one can show that the class of approximations is only slightly larger than the class of Pawlak's rough sets. Therefore, we propose to extend the use of the term *rough set* to all approximations. Our earlier statement may be made clearer now. The main contribution of the paper is a new perspective on rough sets, the notion itself being only slightly modified.

In the paper, we formally define rough sets (in the extended sense) and construct several associated algebraic structures and two ordering relations. One of them, the *knowledge* ordering is especially important and describes the tightness of approximation. It turns out that Pawlak's approximations are best (or maximal) approximations in terms of the knowledge ordering.

In our research we were most strongly motivated by the scenario (2). Consequently, in the paper, we introduce a language of *inclusion-exclusion* that allows us to formulate incomplete specifications of sets based on constraints of the form: "an unknown set contains a given set" or "an unknown set is disjoint with a given set". We use 3-valued Kleene logic to provide semantics to formulas and theories in this language. The connection of rough sets to 3-valued logic is known. It had been noticed early on in [10] and then, more recently, in [2]. This connection is also present implicitly in several other papers on rough sets. Our use of the Kleene logic is, however, novel. We obtain results on the following three key problems underlying a wide range of applications associated with rough sets:

P1: Given a specification T in the language of inclusion-exclusion of a (possibly unknown) set X, what is the tightest rough set approximating X?

P2: How can such approximation be computed?

P3: Given an approximation $\langle L, U \rangle$ of an unknown set, which properties are satisfied by an unknown set X approximated by $\langle L, U \rangle$

The study of problems P1 - P3 is the main focus of our paper.

The paper is organized as follows. The next section provides a brief overview of information systems. The (extended) notion of a rough set and associated algebraic structures are introduced in Section 3. The relationship to Pawlak's definition is also discussed there. The logic of inclusion-exclusion and the problems P1 - P3 are studied in Sections 4 and 5. Some directions for future research are outlined in Section 6.

The study of problems P1 - P3 brings up several interesting computational issues related to the question of existence of short and simple descriptions of sets and their approximations. These questions are related to scenario (3) listed earlier. The area is mostly untouched. Some studies (see, for instance, [18, 3]) point to problems of complexity of descriptions of sets in information systems, thus, potentially, also to complexity of descriptions in terms of rough sets. Yet, no systematic study, to our knowledge, has been undertaken. This and other possible directions for future research are discussed in Section 6.

We hope that our paper offers a new look at rough sets, and an effective and elegant setting in which the theory and applications of rough sets can be investigated.

2. Information systems

In this section, we recall the notion of an information system and describe the corresponding query language [11]. For a detailed treatment of the subject the reader is referred to [11, 15].

An information system is a pair $I = \langle \mathcal{U}, \mathcal{A} \rangle$, where $\mathcal{U}$ is a nonempty set called the *universe* of I and $\mathcal{A} = \{A_1, \ldots, A_n\}$ is a *list* of functions, called *attributes* of I. A function (attribute) $A_i \in \mathcal{A}$ assigns to each element of $\mathcal{U}$ an element from a set D_i called the *domain* of A_i. The list $\mathcal{A}$ of attributes of an information system is often called the *schema* of I.

With each information system I we can associate a function v_I which, to each element $x \in \mathcal{U}$ assigns the description of x in I, that is, the tuple $\langle A_1(x), \ldots, A_n(x) \rangle$. Note that it may be the case (in fact, it is a crucial observation for the theory of rough sets) that different elements $x, y \in \mathcal{U}$ have the *same* description. Informally, it means that our information system is not powerful enough to distinguish between them. For instance, two different individuals may have the same birth date and the same sex and, consequently, will be indistinguishable from the point of an information system based on the schema with these two attributes only. Thus, information system implies an important relation:

$$x \sim_I y \quad \Leftrightarrow \quad v_I(x) = v_I(y).$$

Clearly, the relation $x \sim_I y$ is an equivalence relation that identifies those elements of the universe that have the same description.

To each schema $\mathcal{A}$ we assign now a query language, denoted by $\mathcal{L}_{\mathcal{A}}$. It consists of *terms* built by means of functor symbols $+$ (sum), $\cdot$ (product) and $-$ (negation). The terms of $\mathcal{L}_{\mathcal{A}}$ are defined recursively, as follows:

1. For every attribute $A_i \in \mathcal{A}$ and for every element a in the domain D_i of an attribute A_i, the expression $A_i = a$ is a term
2. If s and t are terms then so are $-s$, $s + t$ and $s \cdot t$.

Terms of the form

$$(A_1 = a_1) \cdot (A_2 = a_2) \cdot \ldots \cdot (A_n = a_n)$$

where $a_i \in D_i$ for every i, $1 \leq i \leq n$, are called *constituent terms* and play an important role in the theory of information systems and rough sets. Clearly, checking if two constituent terms are identical can be accomplished in time proportional to the number of attributes in $\mathcal{A}$.

Terms of $\mathcal{L}_{\mathcal{A}}$ serve as queries to information systems with the schema $\mathcal{A}$. Given an information system $I = \langle \mathcal{U}, \mathcal{A} \rangle$, the *value* of the term (query) t in I, $|t|_I$, is defined recursively by setting:

$$|A_i = a|_I = \{x \in \mathcal{U} : A_i(x) = a\}$$

and by interpreting the product as the set intersection, the sum as the set union, and the negation as the complement with respect to $\mathcal{U}$. Sets assigned to constituent terms are called *constituents*. Let us note that our notion of a constituent is a generalization of a classical set-theoretical notion of a constituent (see [9], Section 1.7). Let us also note that non-empty constituents are precisely the equivalence classes of the relation $\sim_I$. Finally, let us observe that equality of terms in $\mathcal{L}_\mathcal{A}$, interpreted as equality of their values under all information systems with the schema $\mathcal{A}$, can be checked by a propositional prover.

The theory of information systems becomes especially interesting when we adopt one of the *finiteness* conditions:

First finiteness condition: The domains of all attributes are finite

Second finiteness condition: The universe of an information system is finite.

These conditions are independent of each other. It is easy to construct information systems satisfying the first one of them but not the second one and vice versa. Most of the results in the paper hold under *any* of these two conditions. Some, however, rely on the first one and do not, in general, hold under the second one.

In particular, under the first finiteness condition, one can prove the following *normal form* result for terms: for every term $t \in \mathcal{L}_\mathcal{A}$ there is another term, t', such that

1. t' is the sum of constituent terms, and
2. for every information system I, $|t|_I = |t'|_I$.

Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system. The crucial notion for the theory of rough sets is that of a definable set. A subset $X \subseteq \mathcal{U}$ is said to be *definable* in I, if there is a term (query) t such that $|t|_I = X$. In particular, each constituent (the set corresponding to a constituent term) is definable.

Observe that for every two constituents X and X', either $X = X'$ or $X \cap X' = \emptyset$. It is also easy to see that nonempty constituents are minimal nonempty definable sets and that every definable set is a union of, possibly infinitely many, constituents. In fact, one can show that definable sets form a Boolean algebra and that nonempty constituents are its atoms.

Under any of the finiteness conditions a stronger observation holds: any definable set is a union of *finitely* many constituents. Moreover, under the first finiteness condition, there is a bound on the number of terms in such a union that depends only on the schema $\mathcal{A}$ and not on the information system. This observation follows from the normal form result for terms and from the fact that under the first finiteness condition, there are only finitely many constituent terms.

Second finiteness condition has also another related consequence: if every constituent has no more than one element, then every subset of the universe is definable.

It is quite clear that there is a strong database connection. In fact, if we prepend each tuple $v_I(z)$, $z \in \mathcal{U}$, by the unique identifier of z, say $oid(z)$, then the collection of all such extended tuples forms a table that can be viewed as a single class, “flat”, object-oriented database[1]. In addition, the query language described here clearly corresponds to a fragment of SQL: the queries are on a single table, the select clause consists of all attributes (with the exception of the unique identifier attribute *oid*), and range queries, statistical queries and string-matching queries are not permitted. In the paper we will often make references to database intuitions.

[1]We need to use unique identifiers for the elements from the universe since, as mentioned earlier, different elements of the universe of an information system may have the same descriptions and, consequently, would be represented by a single tuple in the database.

3. Approximations and rough sets

In general, not every subset of the universe of an information system $I = \langle \mathcal{U}, \mathcal{A} \rangle$ is definable. In other words, knowledge contained in an information system I is incomplete. Even if a subset of the universe $\mathcal{U}$ is definable, its description may be very complex or we may simply not know it. Therefore, we often have to resort to incomplete or approximate descriptions. In this section we introduce algebraic foundations of the theory of approximations of subsets of the universe of an information system and relate it to the concept of rough sets by Pawlak.

Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system. The key role in our discussion will be played by the boolean algebra of all subsets of $\mathcal{U}$ that are definable in I, that is, can be described by terms of the language $\mathcal{L}_{\mathcal{A}}$. We will denote this algebra by $\mathcal{D}_I$. Under any of the finiteness conditions, the algebra $\mathcal{D}_I$ is a complete boolean algebra.

A straightforward way to approximate a subset X of the universe $\mathcal{U}$ is to provide a lower and an upper bound for it. Since we are interested in approximations that can be expressed in I as values of terms of $\mathcal{L}_{\mathcal{A}}$, we will require that both the lower and the upper bounds be definable in I. Formally, by an *approximation* we mean a pair $\langle L, U \rangle$ such that $L, U \in \mathcal{D}_I$ and $L \subseteq U$. Each such pair $\langle L, U \rangle$ can be viewed as an approximation of any set $Z \subseteq \mathcal{U}$ (definable in I or not) such that $L \subseteq Z \subseteq U$. An approximation is not the same notion as that of a rough set as defined by Pawlak. But both concepts are very closely related (we will introduce Pawlak's rough sets and discuss this relationship later in this section). Thus, somewhat abusing the terminology, throughout the paper we refer to approximations as *rough sets*.

We denote the collection of all rough sets (approximations) in an information system I by $\mathcal{R}_I$. This structure can be endowed with an ordering called the *knowledge ordering*. It is denoted by $\preceq_{kn}$, and is defined as follows:

$$\langle L_1, U_1 \rangle \preceq_{kn} \langle L_2, U_2 \rangle \text{ if } L_1 \subseteq L_2 \text{ and } U_2 \subseteq U_1.$$

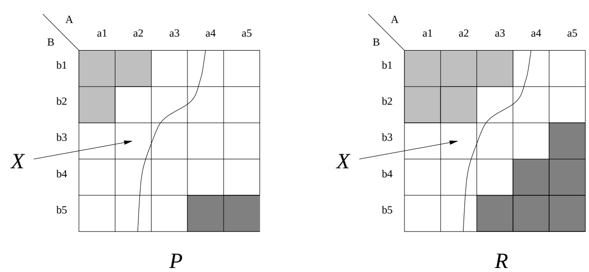

Figure 1. Two rough sets in the relation $\preceq_{kn}$

Figure 1 presents two rough sets $P = \langle L_1, U_1 \rangle$ and $R = \langle L_2, U_2 \rangle$. The lower approximations L_1 and L_2 are shown as lightly shaded. Complements of the upper approximations U_1 and U_2 are darkly shaded. These sets are defined by the following terms:

L_1: $((A = a1) \cdot (B = b1)) + ((A = a2) \cdot (B = b1)) + ((A = a1) \cdot (B = b2))$
U_1: $-(((A = a4) + (A = a5)) \cdot (B = b5))$

L_2: $(((A = a1) + (A = a2)) \cdot ((B = b1) + (B = b2)))+$
$((A = a3) \cdot (B = b1))$

U_2: $(A = a1) + (A = a2) + (B = b1) + (B = B2)+$
$((A = a3) \cdot (B = b3)) + (A = a3) \cdot (B = b4))+$
$(A = a4) \cdot (B = b3))$

Clearly, $P \preceq_{kn} R$.

The knowledge ordering is crucial for our considerations and requires some explanation. If pairs $\langle L_1, U_1\rangle$ and $\langle L_2, U_2\rangle$ are *approximations* and $\langle L_1, U_1\rangle \preceq_{kn} \langle L_2, U_2\rangle$ then the pair $\langle L_2, U_2\rangle$ is a *tighter* approximation (contains more precise knowledge about an unknown set Z that both pairs approximate). In particular, the set X of elements "to the left" of the curved line in Figure 1 is approximated both by P and by R. It is clear that R is a tighter approximation, that is, provides more knowledge about the set X. This intuition motivates the use of the term *knowledge* in reference to the ordering $\preceq_{kn}$.

If L is not a subset of U, the pair $\langle L, U\rangle$ cannot be interpreted as an approximation (unless we want to interpret all such pairs as *inconsistent* approximations). Still, the ordering $\preceq_{kn}$ can be extended to the whole cartesian product $\mathcal{D}_I \times \mathcal{D}_I$ and, in fact, also to the cartesian product $\mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U})$. We will consider these two structures, too, since they simplify some of the technical arguments later in the paper. The following result gathers the most important properties of sets $\mathcal{R}_I$, $\mathcal{D}_I \times \mathcal{D}_I$, $\mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U})$ and the ordering $\preceq_{kn}$.

Proposition 3.1. *For every set $\mathcal{U}$, $\langle \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U}), \preceq_{kn}\rangle$ is a complete lattice. For every information system I satisfying any of the finiteness conditions, the structure $\langle \mathcal{D}_I \times \mathcal{D}_I, \preceq_{kn}\rangle$ is a complete lattice and $\langle \mathcal{R}_I, \preceq_{kn}\rangle$ is a complete lower semi-lattice.*

Let us note that $\langle \emptyset, \mathcal{U}\rangle$ is the least and $\langle \mathcal{U}, \emptyset\rangle$ is the greatest element of $\langle \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U}), \preceq_{kn}\rangle$ and of $\langle \mathcal{D}_I \times \mathcal{D}_I, \preceq_{kn}\rangle$. The pair $\langle \emptyset, \mathcal{U}\rangle$ is also the least element of the poset $\langle \mathcal{R}_I, \preceq_{kn}\rangle$. The maximal elements in $\langle \mathcal{R}_I, \preceq_{kn}\rangle$ are pairs $\langle X, X\rangle$, where $X \in \mathcal{D}_I$.

Proposition 3.1 allows us to derive properties of rough sets. Most importantly, it allows us to apply the theorem by Knaster and Tarski [21] on existence of fixpoints of monotone operators on complete lattices.

The sets $\langle \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U})\rangle$, $\langle \mathcal{D}_I \times \mathcal{D}_I\rangle$ and $\mathcal{R}_I$ (in fact, any collection of pairs of sets) can also be ordered by the so-called *inclusion ordering* $\preceq_{in}$. It is defined as follows:

$$\langle L_1, U_1\rangle \preceq_{in} \langle L_2, U_2\rangle \text{ if } L_1 \subseteq L_2 \text{ and } U_1 \subseteq U_2.$$

It is easy to see that all three sets are complete lattices under the ordering $\preceq_{in}$ (in the case of $\mathcal{D}_I \times \mathcal{D}_I$ and $\mathcal{R}_I$ we need to assume one of the finiteness conditions).

It is not clear whether the ordering $\preceq_{in}$ plays any major role in the theory of rough sets. However, let us note that the structures $\langle \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U}), \preceq_{kn}, \preceq_{in}\rangle$ and $\langle \mathcal{D}_I \times \mathcal{D}_I, \preceq_{kn}, \preceq_{in}\rangle$ (this latter one under any of the finiteness conditions) form complete bilattices [6, 5].

We will now discuss connections between the concept of a rough set as defined above and the original one introduced by Pawlak [14]. Pawlak observed that when an information system

$I = \langle \mathcal{U}, \mathcal{A} \rangle$ satisfies any of the finiteness conditions then, for every set $X \subseteq \mathcal{U}$, there exists a greatest definable set X' such that $X' \subseteq X$ and, similarly, there exists a smallest definable set X'' such that $X \subseteq X''$. These sets are denoted by $\underline{X}$ and $\overline{X}$, respectively, and called *lower* and *upper approximations* of X. It is necessary to adopt at least one finiteness condition as, in general, there are information systems in which, for some subsets X of the universe, the lower or the upper approximations (or both) are not defined. Pawlak called pairs of the form $\langle \underline{X}, \overline{X} \rangle$, where $X \subseteq \mathcal{U}$, *rough sets.*

If $\langle L, U \rangle$ is a rough set, and X is a subset of $\mathcal{U}$, then we say that X is *dense* in $\langle L, U \rangle$ if $\underline{X} = L$ and $\overline{X} = U$. In such case, we also say that the rough set $\langle L, U \rangle$ is *concrete.* Thus, Pawlak's rough sets are precisely those rough set according to our definition that are concrete. In Figure 2, we present the concrete rough set corresponding to the set X of elements "to the left" of the curved line (light shade indicates the lower approximation, dark shade indicates the complement of the upper approximation). The set X was also discussed in the context of Figure 1. Clearly, $P \preceq_{kn} R \preceq_{kn} S$ (P and R are as in Figure 1). In fact, S is the $\preceq_{kn}$-largest approximation to the set X.

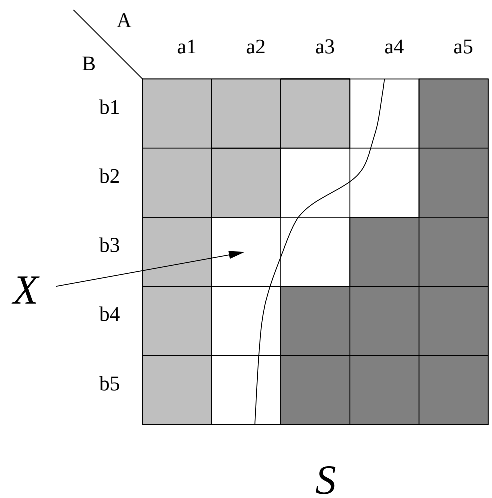

Figure 2. A subset X of $\mathcal{U}$ dense in a rough set S

It follows directly from the definition that $\underline{X} \subseteq \overline{X}$ and that both approximations are definable. Thus, Pawlak's rough sets are rough sets in our sense, as well. In general, the converse does not hold. However, the connection is very strong, as explained in the next result.

Proposition 3.2. *A rough set $\langle L, U \rangle$ is of the form $\langle \underline{X}, \overline{X} \rangle$, for some set $X \subseteq \mathcal{U}$ if and only if for every $x \in U \setminus L$, the constituent of x has at least two elements.*

Proposition 3.2 implies immediately that if every constituent of an information system has at least two elements, the class of rough sets according to the definition by Pawlak coincides with the class $\mathcal{R}_I$ of rough sets as defined in this paper. Thus, both concepts are very closely related which justifies our use of the term. Let us stress again that the main contribution of our work is not in the change of the definition but in the change of perspective. A rough set as defined by Pawlak is intimately connected to the underlying subset of the universe that determines it. This set is, however, usually unknown. Starting with the notion of an approximation, not tied to any subset of the universe in particular, seems to be more natural. It leads directly to orderings $\preceq_{kn}$ and $\preceq_{in}$ and allows us to exploit algebraic techniques in our study of approximations.

We conclude this section by discussing some simple properties of rough sets. Our first result states that Pawlak's rough sets provide the best approximations.

Proposition 3.3. *Let I satisfy one of the finiteness conditions. Then, for every set $X \subseteq \mathcal{U}$, $\langle \underline{X}, \overline{X} \rangle$ is the $\preceq_{kn}$-greatest rough set approximating X.*

Proof: If $R = \langle L, U \rangle$ approximates X, then $L \subseteq X$. Since L is definable, $L \subseteq \underline{X}$. Similarly, $\overline{X} \subseteq U$. Thus $R \preceq_{kn} \langle \underline{X}, \overline{X} \rangle$. □

The next result, due to Pawlak [14], deals with the ordering $\preceq_{in}$. It says that as sets grow, so do, with respect to $\preceq_{in}$, their Pawlak's approximations.

Proposition 3.4. *If $X \subseteq Y$ then $\langle \underline{X}, \overline{X} \rangle \preceq_{in} \langle \underline{Y}, \overline{Y} \rangle$.*

Finally, let us observe that unknown sets (concepts) are often, especially in learning, specified by positive and negative examples, that is, two finite and disjoint sets of elements (subsets of the universe of an information system I): those that are *in* and those that are *out*. We will call such a pair of sets a *sample*.

Consider a sample $\langle P, N \rangle$. We say that an information system I is *adequate* for $\langle P, N \rangle$ if for no elements $x \in P$ and $y \in N$ we have $x \sim_I y$. Informally, I is adequate for a sample $\langle P, N \rangle$ if it allows us to distinguish between positive and negative examples of the set (concept) that we attempt to describe.

In general, samples provide only an incomplete description of a set. Therefore, we will be interested in approximations (rough sets) that can be associated with (learned from) a sample. We say that a rough set $\langle L, U \rangle$ is *consistent with a sample* $\langle P, N \rangle$ if $P \subseteq L$, $N \cap U = \emptyset$. An information system I is *consistent with a sample* $\langle P, N \rangle$ if there is a rough set over I consistent with $\langle P, N \rangle$. We have the following simple result.

Proposition 3.5. *Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system. Then:*

1. *I is consistent with $\langle P, N \rangle$ if and only if I is adequate for $\langle P, N \rangle$*
2. *If I is consistent with $\langle P, N \rangle$ then there is a $\preceq_{kn}$-least rough set R consistent with $\langle P, N \rangle$.*

Proof: (1) Let $\langle L, U \rangle$ be a rough set over I consistent with $\langle P, N \rangle$. Consider $x \in P$ and $y \in N$. Then, $x \in L$ and $y \in \mathcal{U} \setminus U$. Since both L and $\mathcal{U} \setminus U$ are definable and disjoint, x and y are not equivalent with respect to $\sim_I$. Thus, I is adequate for $\langle P, N \rangle$.

Conversely, assume that I is adequate for $\langle P, N \rangle$. Let $P = \{x_1, \ldots, x_m\}$ and $N = \{y_1, \ldots, y_n\}$. For $1 \leq i \leq m$, let t_i be the constituent term such that $x_i \in |t_i|_I$ and, for $1 \leq j \leq n$, let s_j be the constituent term such that $y_j \in |s_j|_I$. The terms t_i, s_j are well-defined, as each element of $\mathcal{U}$ belongs to some constituent set. By the assumption of adequacy, $t_i \neq s_j$, for all i, j. Define

$t = t_1 + \ldots + t_m$ and $s = -(s_1 + \ldots + s_n)$. Put $L = |t|_I$ and $U = |s|_I$. Then, by the remarks above, $\langle L, U\rangle$ is consistent with $\langle P, N\rangle$.
(2) It can be shown that the rough set constructed in the second part of the proof of (1) is the $\preceq_{kn}$-least rough set consistent with a sample $\langle P, N\rangle$. $\square$

We will denote the rough set constructed in the proof of Proposition 3.5 by $R(P, N)$. It encodes the entire knowledge (with respect to the underlying information system I) carried by the sample $\langle P, N\rangle$. Namely, it approximates every definable in I set X such that $P \subseteq X$ and $N \cap X = \emptyset$ (there is a close similarity here with the notion of version space in learning [12]).

In the context of rough sets (and 3-valued logic) we can extend our discussion to the case when I is not adequate for the sample $\langle P, N\rangle$. In such case, there is no set X definable in I and such that $P \subseteq X$ and $N \cap X = \emptyset$. That is, P and N cannot be "separated" in I. But they can be separated "as much as possible". Given a sample $\langle P, N\rangle$, let us call any sample $\langle P', N'\rangle$ such that $P' \subseteq P$ and $N' \subseteq N'$ a *subsample* of $\langle P, N\rangle$.

Proposition 3.6. *Let I be an information system and let $\langle P, N\rangle$ be a sample. Then, there is a $\preceq_{in}$-largest subsample of $\langle P, N\rangle$ with which I is consistent.*

Proof: Clearly, $\langle \emptyset, \emptyset\rangle$ is a subsample of $\langle P, N\rangle$ consistent with I. Moreover, it is easy to see that if subsamples (P', N') and (P'', N'') of (P, N) are consistent with I, the subsample $(P' \cup P'', N' \cup N'')$ of (P, N) is also consistent with I. Thus, the assertion follows. $\square$

Let us denote this $\preceq_{in}$-largest subsample of (P, N), guaranteed by Proposition 3.6, by (P^I, N^I). The rough set $R(P^I, N^I)$, guaranteed by Proposition 3.5, describes all those definable sets in I that separate P^I from N^I or, speaking informally, separate as much of P from N as possible.

We will now find an alternative characterization of the rough set $R(P^I, N^I)$. To this end, let us call a rough set $\langle L, U\rangle$ *weakly consistent* with $\langle P, N\rangle$, if $N \cap L = \emptyset$, and $P \cap (\mathcal{U} \setminus U) = \emptyset$. Rough sets that are weakly consistent with a sample $\langle P, N\rangle$ always exist. For instance, $\langle \emptyset, \mathcal{U}\rangle$ is one such set. Moreover, under any of the finiteness conditions, every $\preceq_{kn}$-chain consisting of rough sets weakly consistent with $\langle P, N\rangle$ has a least upper bound. It is also easy to see that this least upper bound is itself weakly consistent with $\langle P, N\rangle$. Thus, for every rough set $\langle L, U\rangle$ weakly consistent with $\langle P, N\rangle$, there exists a $\preceq_{kn}$-maximal rough set $\langle L^m, U^m\rangle$ weakly consistent with $\langle P, N\rangle$ and such that $\langle L, U\rangle \preceq_{kn} \langle L^m, U^m\rangle$.

We then have the following property.

Theorem 3.1. *Let I satisfies any finiteness condition and let $\langle P, N\rangle$ be a sample. Then, $R(P^I, N^I)$ is the greatest lower bound of all maximal rough sets that are weakly consistent with $\langle P, N\rangle$.*

Proof: Assume that $R(P^I, N^I) = \langle R_l, R_u\rangle$. Consider an arbitrary maximal rough set weakly consistent with $\langle P, N\rangle$, say $\langle L, U\rangle$. Let $x \in R_l$. Define t_x to be a constituent term of x and let $C_x = |t_x|_I$.

Assume that $C_x \cap U \neq \emptyset$. Since U is definable, $C_x \subseteq U$. By the definition of R_l, there is $p \in P$ such that $p \in R_l$ and $x \sim_I p$. Thus, $p \in U$, a contradiction (as $\langle L, U\rangle$ is weakly consistent with $\langle P, N\rangle$). It follows then that $C_x \cap U = \emptyset$. Consequently, $\langle L \cup C_x, U\rangle$ is a rough set.

Assume that for some element $n \in N$, $n \in L \cup C_x$. Since $\langle L, U\rangle$ is weakly consistent with $\langle P, N\rangle$, $n \in C_x$. It follows that $n \sim_I x$. Consequently, $n \sim_I p$ and, thus, $n \in R_l$. This is a contradiction as $R(P^I, N^I)$ is weakly consistent with $\langle P, N\rangle$.

Thus, $\langle L \cup C_x, U\rangle$ is a weakly consistent rough set with $\langle P, N\rangle$. Since $C_x \neq \emptyset$ and since $\langle L, U\rangle$ is a maximal rough set weakly consistent with $\langle P, N\rangle$ it follows that $C_x \subseteq L$. In particular $x \in L$. Hence, $R_l \subseteq L$. Similarly, one can show that $U \subseteq R_u$. Consequently, $R(P^I, N^I) \preceq_{kn} \langle L, U\rangle$.

Consider now a rough set $\langle L_0, U_0\rangle$ such that $\langle L_0, U_0\rangle \preceq_{kn} \langle L, U\rangle$ for every $\langle L, U\rangle$ that is a maximal rough set weakly consistent with $\langle P, N\rangle$. Let $x \in L_0$. Assume that $x \notin R_l$. As before, define t_x to be a constituent term of x and let $C_x = |t_x|_I$. By the definition of R_l, there are two possibilities: (1) $C_x \cap (P \cup N) = \emptyset$, and (2) $C_x \cap N \neq \emptyset$.

Since L_0 is definable, it follows that $C_x \subseteq L_0$. Let $\langle L, U\rangle$ be a maximal rough set weakly consistent with $\langle P, N\rangle$ (as we observed earlier, any of the finiteness conditions implies that such maximal sets exist). Then, $\langle L_0, U_0\rangle \preceq_{kn} \langle L, U\rangle$ and so $C_x \subseteq L$. In the case when (2) holds, we get an immediate contradiction with weak consistency of $\langle L, U\rangle$. So, assume that (1) holds. It is clear that the rough set $\langle L \setminus C_x, U \cup C_x\rangle$ is also a rough set weakly consistent with $\langle P, N\rangle$. Consequently, there is a maximal rough set $\langle L', U'\rangle$ weakly consistent with $\langle P, N\rangle$ and such that $C_x \cap L' = \emptyset$. Thus, $\langle L_0, U_0\rangle \not\preceq_{kn} \langle L', U'\rangle$, a contradiction. It follows that $x \in R_l$ and that $L_0 \subseteq R_l$. in a similar way, one can prove that $R_u \subseteq U_0$. Thus, $\langle L_0, U_0\rangle \preceq_{kn} R(P^I, N^I)$, and the assertion follows. □

Finally, let us observe that when an information system is inadequate for a sample $\langle P, N\rangle$, adding new attributes to the language yields information systems allowing for more complete separation of positive and negative elements. We have the following straightforward result.

Proposition 3.7. *Let $I = \langle \mathcal{U}, \mathcal{A}\rangle$ and $J = \langle \mathcal{U}, \mathcal{A}'\rangle$ be information systems. If $\mathcal{A} \subseteq \mathcal{A}'$ then for every sample $\langle P, N\rangle$, $\langle P^I, N^I\rangle \preceq_{in} \langle P^J, N^J\rangle$.*

4. Logic of inclusion-exclusion

Pawlak's rough sets and rough sets introduced here are motivated by the need to reason about unknown sets of records — sets for which we have only an incomplete specification. We will now investigate this main application of rough sets in more detail.

It is often the case that a set of interest is unknown but some information about it is available. For instance, we may know about some sets being contained in it and some other sets being disjoint with it. We will introduce a language to describe constraints of these types.

Given the schema $\mathcal{A}$ of an information system and the corresponding language $\mathcal{L}_{\mathcal{A}}$ we define the *language of inclusion-exclusion* for $\mathcal{A}$, $\mathcal{L}^{ie}_{\mathcal{A}}$ as follows. The atoms of $\mathcal{L}^{ie}_{\mathcal{A}}$ are expressions of the form $\mathbf{in}(t)$ and $\mathbf{ex}(t)$, where $t \in \mathcal{L}_{\mathcal{A}}$. Next, if φ_1 and φ_2 are formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$ then so are $\varphi_1 \wedge \varphi_2$, $\varphi_1 \vee \varphi_2$, $\varphi_1 \Rightarrow \varphi_2$ and $\neg\varphi_1$.

Intuitively, a formula $\mathbf{in}(t)$ describes the constraint that an unknown subset of the universe of an information system I contains the answer to the query t, that is, the set $|t|_I$. Similarly, a formula $\mathbf{in}(r) \Rightarrow \mathbf{in}(s) \vee \mathbf{ex}(t)$ describes the constraint that if a set contains $|r|_I$ then it contains $|s|_I$ or is disjoint with $|t|_I$. We will now make this intuition precise by defining the satisfiability relation between subsets of the universe of an information system and formulas in the language $\mathcal{L}^{ie}_{\mathcal{A}}$.

Given an information system $I = \langle \mathcal{U}, \mathcal{A} \rangle$ and a set $X \subseteq \mathcal{U}$ (X may but does not have to be definable) we define

$$[\mathbf{in}(t)]_X = \begin{cases} \mathbf{1} & \text{if } |t|_I \subseteq X \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Similarly we define

$$[\mathbf{ex}(t)]_X = \begin{cases} \mathbf{1} & \text{if } |t|_I \cap X = \emptyset \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Next, we extend the definition of $[\varphi]_X$ to all formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$ interpreting $\neg$, $\wedge$, $\vee$ and $\Rightarrow$ in a standard way in the boolean algebra of logical values. That is, $[\neg\varphi]_X = 1 - [\varphi]_X$, and $[\varphi \wedge \psi]_X = \min([\varphi]_X, [\psi]_X)$, etc. We say that $X \models_I \varphi$ if $[\varphi]_I = 1$. When T is a theory, that is, a set of formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$, we say that X is a *model* of T (or that X *satisfies* T) if $X \models_I \varphi$ for all $\varphi \in T$. We will denote it by $X \models_I T$ and define $Mod_I(T) = \{X \subseteq \mathcal{U} : X \models_I T\}$.

We say that a theory T in the language $\mathcal{L}^{ie}_{\mathcal{A}}$ is *consistent with* $I = \langle \mathcal{U}, \mathcal{A} \rangle$ if there is a set $X \subseteq \mathcal{U}$ such that $X \models_I T$.

Before we proceed to the main questions listed in the introduction, let us note some simple but interesting properties of the 2-valued semantics of the language of inclusion-exclusion. Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system. We say that a set $X \subseteq \mathcal{U}$ is *constituent-complete* (with respect to I) if for every constituent term t, $X \models_I \mathbf{in}(t)$ or $X \models_I \mathbf{ex}(t)$. We now have the following characterization of definable sets.

Proposition 4.1. *If an information system $I = \langle \mathcal{U}, \mathcal{A} \rangle$ satisfies the first finiteness assumption, then a set $X \subseteq \mathcal{U}$ is definable in I if and only if X is constituent-complete.*

The language of inclusion-exclusion can distinguish between empty constituents, 1-element constituents and constituents with more than one element. It does not, in general, distinguish between cardinalities greater than or equal to 2.

Proposition 4.2. *Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system and let $X \subseteq \mathcal{U}$.*

1. *$|t|_I = \emptyset$ if and only if for every set $X \subseteq \mathcal{U}$, $X \models_I \mathbf{in}(t) \wedge \mathbf{ex}(t)$.*
2. *$|t|_I$ is a one-element set if and only if for every set $X \subseteq \mathcal{U}$, $X \models_I \neg\mathbf{in}(t) \Leftrightarrow \mathbf{ex}(t)$.*
3. *If t is a constituent such that $|t|_I \geq 2$ then for any $k \geq 2$ there is an information system $I' = \langle \mathcal{U}', \mathcal{A} \rangle$ (notice that the set of attributes is the same as in I) such that:*

    (a) $|t|_{I'} = k$

(b) *For every $X \subseteq \mathcal{U}$ there is $X' \subseteq \mathcal{U}'$ such that X and X' satisfy precisely the same formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$.*

Next, let us note that sets that are indistinguishable in an information system I satisfy precisely the same formulas from $\mathcal{L}^{ie}_{\mathcal{A}}$. Recall that a set $X \subseteq \mathcal{U}$ is dense in a rough set $\langle L, U\rangle$ if $\underline{X} = L$ and $\overline{X} = U$ (that is, if $\langle L, U\rangle$ is the rough set of X in the sense of Pawlak). Sets X and Y are *indistinguishable* if they are dense in the same rough set (that is, if both have the same rough set in the sense of Pawlak).

Theorem 4.1. (Indistinguishability theorem) *Let I satisfy one of the finiteness conditions. Then, for $X, Y \subseteq \mathcal{U}$, X and Y are indistinguishable if and only if for every formula φ of the language of inclusion-exclusion*

$$X \models_I \varphi \Leftrightarrow Y \models_I \varphi.$$

Theorem 4.1 demonstrates that rough sets are, really, about indistinguishability in the language $\mathcal{L}^{ie}_{\mathcal{A}}$. Thus, any strengthening of the concept of a rough set (for instance so we would be able to formally express the quality of approximation) requires strengthening of the language of inclusion-exclusion.

We will now formally state and study general problems that arise in the context of reasoning about properties of unknown sets specified by means of formulas from $\mathcal{L}^{ie}_{\mathcal{A}}$. First, given an information system $I = \langle \mathcal{U}, \mathcal{A}\rangle$ and a theory from $\mathcal{L}^{ie}_{\mathcal{A}}$ describing available information about an unknown set $X \subseteq \mathcal{U}$, the question is to determine, as accurately as possible, the extent of X (problem P1). Next, there is a question of computing this tightest approximation (problem P2). Finally, given a rough set that approximates an unknown set $X \subseteq \mathcal{U}$, the question is to establish properties (expressed as formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$) that X has (problem P3). The study of these questions is the main goal for the remainder of the paper.

We start with the first problem. We will show that given a theory T in the language $\mathcal{L}^{ie}_{\mathcal{A}}$, there exists a rough set providing the best approximation to all sets X that satisfy T.

Indeed, let T be a consistent theory in the language of inclusion-exclusion $\mathcal{L}^{ie}_{\mathcal{A}}$. Let A_T be the class of all rough sets $\langle L, U\rangle$ such that for every $X \in Mod_I(T)$, $L \subseteq X \subseteq U$ (that is, A_T consists of all rough sets $\langle L, U\rangle$ such that $\langle L, U\rangle \preceq_{kn} \langle X, X\rangle$ whenever $X \models_I T$). Then, clearly, A_T is nonempty — $\langle \emptyset, \mathcal{U}\rangle \in A_T$. Now, we can prove the following fact.

Theorem 4.2. (Approximation theorem) *Assume I satisfies one of the finiteness conditions. Let T be a consistent theory in the language of inclusion-exclusion. Then A_T possesses a $\preceq_{kn}$-greatest element. That is, there exists the $\preceq_{kn}$-greatest rough set $\langle L, U\rangle$ such that if $X \subseteq \mathcal{U}$ and $X \models_I T$ then $L \subseteq X$, and $X \subseteq U$.*

Proof: Let X be a model of T (such a model exists since T is consistent). Then, the class A_T is nonempty. Moreover, the class A_T is closed under finite joins. That is, for every $\langle L, U\rangle, \langle L', U'\rangle \in A_T$, $\langle L \cup L', U \cap U'\rangle \in A_T$. Indeed, let $\langle L, U\rangle, \langle L', U'\rangle \in A_T$. Then, $\langle L, U\rangle \preceq_{kn} \langle X, X\rangle$ and $\langle L', U'\rangle \preceq_{kn} \langle X, X\rangle$. Consequently, $\langle L \cup L', U \cap U'\rangle \preceq_{kn} \langle X, X\rangle$. Since X is an arbitrary model of T, $\langle L \cup L', U \cap U' \rangle \in A_T$. By a finiteness condition, the class A_T is finite. Hence, it contains the join of all its elements and this element is the $\preceq_{kn}$-greatest element of A_T. $\square$

Earlier we used notation $\langle \underline{X}, \overline{X} \rangle$, where $X \subseteq \mathcal{U}$, to denote lower and upper approximations to a set X (or, equivalently, a concrete rough sets determined by X). The rough set $\langle \underline{X}, \overline{X} \rangle$ is the $\preceq_{kn}$-greatest approximation to X. Given a theory T, by $\langle \underline{T}, \overline{T} \rangle$ we denote the $\preceq_{kn}$-greatest element of A_T, whose existence is guaranteed by Theorem 4.2. Since $\langle \underline{T}, \overline{T} \rangle$ is the $\preceq_{kn}$-greatest element of A_T, if $x \notin \underline{T}$, then there is X satisfying T such that $x \notin X$. Similarly, if $x \notin \overline{T}$, then there is X satisfying T such that $x \in X$. Thus, $\langle \underline{T}, \overline{T} \rangle$ is the best approximation of an unknown set specified by T, if T is all we know about it, justifying extending the notation $\langle \underline{\cdot}, \overline{\cdot} \rangle$ to the case of theories in $\mathcal{L}^{ie}_{\mathcal{A}}$.

Theorem 4.2 asserts only the existence of the set $\langle \underline{T}, \overline{T} \rangle$. It does not imply a method to construct it (note that our proof of Theorem 4.2 relies on the knowledge of the family of A_T of all possible sets that could be represented by T). In the next section we will develop tools that will allow us to tackle the second problem listed earlier and, in addition, will yield techniques to construct the approximation $\langle \underline{T}, \overline{T} \rangle$ for some special classes of theories T.

5. Three-valued logic of inclusion-exclusion

In order to further study the problems stated in the previous section we need to introduce a 3-valued semantics for theories in the language $\mathcal{L}^{ie}_{\mathcal{A}}$. We use the 3-valued logic of Kleene and introduce the 3-valued satisfiability relation between rough sets and formulas from $\mathcal{L}^{ie}_{\mathcal{A}}$ in a similar way as the 2-valued satisfaction relation was introduced in Section 4.

Kleene 3-valued logic, [8], pp. 332-335, is based on three logical values, $\mathbf{1}$, $\mathbf{0}$, and $\mathbf{u}$. These logical values are ordered by a relation $\leq_{tr}$ (often referred to as the *truth ordering*) $\mathbf{0} \leq_{tr} \mathbf{u} \leq_{tr} \mathbf{1}$. The operations $\wedge$ and $\vee$ on the truth values $\mathbf{1}$, $\mathbf{0}$, and $\mathbf{u}$ are defined as meet and join with respect to relation $\leq_{tr}$. The complement operation, $(\cdot)^{-1}$, is defined as follows:

$$\mathbf{0}^{-1} = \mathbf{1}, \quad \mathbf{1}^{-1} = \mathbf{0}, \quad \mathbf{u}^{-1} = \mathbf{u}.$$

The truth values in the Kleene logic are also ordered by another ordering, the knowledge ordering, $\leq_{kn}$ in which $\mathbf{u}$ is the least element and $\mathbf{1}$, $\mathbf{0}$ are the maximal elements.

We will now define a 3-valued satisfiability relation. Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system and let $\langle L, U \rangle \in \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U})$ be a pair of subsets of $\mathcal{U}$. We first define

$$[\mathbf{in}(t)]_{\langle L,U \rangle} = \begin{cases} \mathbf{1} & \text{if } |t|_I \subseteq L \\ \mathbf{0} & \text{if } |t|_I \setminus U \neq \emptyset \\ \mathbf{u} & \text{otherwise} \end{cases}$$

and

$$[\mathbf{ex}(t)]_{\langle L,U \rangle} = \begin{cases} \mathbf{1} & \text{if } |t|_I \cap U = \emptyset \\ \mathbf{0} & \text{if } |t|_I \cap L \neq \emptyset \\ \mathbf{u} & \text{otherwise} \end{cases}$$

Next we extend the definition of $[\varphi]_{\langle L,U\rangle}$ to all formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$. We interpret $\neg$, $\wedge$ and $\vee$ as the Kleene complement, meet and join. The interpretation of $\Rightarrow$ is implied by the fact that $p \Rightarrow q$ is equivalent, in Kleene's logic, to $\neg p \vee q$. Finally, as in Section 4, we define

$$\langle L,U\rangle \models_{I,3} \varphi \text{ if } [\varphi]_{\langle L,U\rangle} = \mathbf{1}.$$

The notions of 2-valued and 3-valued satisfiability are closely related. First, for complete rough sets, that is, for rough sets of the form $\langle X,X\rangle$ they coincide.

Proposition 5.1. *Let X be a definable set in I and let $\varphi \in \mathcal{L}^{ie}_{\mathcal{A}}$. Then $X \models_I \varphi$ if and only if $\langle X,X\rangle \models_{I,3} \varphi$.*

Moreover, the relation $\models_{I,3}$ approximates $\models_I$ for dense sets.

Theorem 5.1. *Let $R = \langle L,U\rangle$ be a rough set and let $\varphi \in \mathcal{L}^{ie}_{\mathcal{A}}$. If $R \models_{I,3} \varphi$ then for every X such that X is dense in R (that is $\underline{X} = L$ and $\overline{X} = U$) we have $X \models_I \varphi$.*

Theorem 5.1 tells us that the satisfaction relation for a rough set R (defined by means of 3-valued logic) truly approximates 2-valued satisfaction relation for all subsets X of $\mathcal{U}$ that are dense in R.

We now resume our study of the three main problems listed in the introduction. We have the following key property connecting the satisfaction relation $\models_{I,3}$ with the knowledge ordering.

Theorem 5.2. *Let $R_1 = \langle L,U\rangle$ and $R_2 = \langle L',U'\rangle$ be two elements of $\mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U})$ such that $R_1 \preceq_{kn} R_2$. Let $\varphi \in \mathcal{L}^{ie}_{\mathcal{A}}$. Then, $[\varphi]_{R_1} \leq_{kn} [\varphi]_{R_2}$. In particular, if $R_1 \preceq_{kn} R_2$ and $R_1 \models_{I,3} \varphi$, then $R_2 \models_{I,3} \varphi$.*

Proof: We prove only the first assertion. The second one is its immediate consequence. We proceed by induction on the complexity of the formula φ. First, let $\varphi = \mathbf{in}(t)$. If $[\varphi]_{\langle L,U\rangle} = \mathbf{1}$ then $|t|_I \subseteq L$. From the assumption $\langle L,U\rangle \preceq_{kn} \langle L',U'\rangle$ it follows that $L \subseteq L'$ so $|t|_I \subseteq L'$. Thus $[\varphi]_{\langle L',U'\rangle} = \mathbf{1}$. If $[\varphi]_{\langle L,U\rangle} = \mathbf{0}$ then $|t|_I \cap U = \emptyset$. But $U' \subseteq U$, and so $|t|_I \cap U' = \emptyset$. Thus $[\varphi]_{\langle L',U'\rangle} = \mathbf{0}$. When $[\varphi]_{\langle L,U\rangle} = \mathbf{u}$ then there is nothing to prove since $\mathbf{u}$ is the least element of the ordering $\leq_{kn}$. The argument for the case of $\varphi = \mathbf{ex}(t)$ is similar.

In the inductive step, three cases need to be considered. If $[\neg\varphi]_{\langle L,U\rangle} = \mathbf{0}$ then $[\varphi]_{\langle L,U\rangle} = \mathbf{1}$. By the inductive assumption, $[\varphi]_{\langle L',U'\rangle} = \mathbf{1}$, and so $[\neg\varphi]_{\langle L',U'\rangle} = \mathbf{0}$. The case of $[\varphi]_{\langle L,U\rangle} = \mathbf{0}$ is similar. In the case $[\neg\varphi]_{\langle L,U\rangle} = \mathbf{u}$, there is nothing to prove as $\mathbf{u}$ is the least element. If $\varphi = \varphi_1 \vee \varphi_2$ and $[\varphi]_{\langle L,U\rangle} = \mathbf{1}$ then $[\varphi_1]_{\langle L,U\rangle} = \mathbf{1}$ or $[\varphi_2]_{\langle L,U\rangle} = \mathbf{1}$. By inductive assumption $[\varphi_1]_{\langle L',U'\rangle} = \mathbf{1}$ or $[\varphi_2]_{\langle L',U'\rangle} = \mathbf{1}$, thus $[\varphi]_{\langle L',U'\rangle} = \mathbf{1}$. The case of $[\varphi]_{\langle L,U\rangle} = \mathbf{0}$ is similar and the case of $\mathbf{u}$ can be dealt with as before. The case of conjunction is similar to the case of disjunction. $\square$

Theorem 5.2 provides an additional justification for the term *knowledge ordering* used in reference to the ordering $\preceq_{kn}$. Namely, as approximations get more precise (grow with the knowledge ordering), our knowledge about formulas from $\mathcal{L}^{ie}_{\mathcal{A}}$ grows, too.

Theorem 5.2 has a corollary that provides an answer to the problem P3 listed in the introduction. It allows us to draw conclusions about properties of unknown sets based on the properties of their approximations.

Corollary 5.1. *Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system and let X be a subset of $\mathcal{U}$. Let R be a rough set that approximates X, that is, $R \preceq_{kn} \langle X, X \rangle$. Then, for every $\varphi \in \mathcal{L}^{ie}_{\mathcal{A}}$, if $R \models_{I,3} \varphi$ then $X \models_I \varphi$.*

Proof: Since $R \preceq_{kn} \langle X, X \rangle$, it follows by Theorem 5.2 that if $R \models_{I,3} \varphi$ then $\langle X, X \rangle \models_{I,3} \varphi$. But for complete rough sets, the relation $\models_{I,3}$ coincides with $\models_I$ (Proposition 5.1). $\square$

Corollary 5.1 implies that if we are given an approximation R of an unknown set X then all properties satisfied by R (in 3-valued logic) are also satisfied by X (in 2-valued logic).

We return now to the question left open at the end of the previous section: how to compute the best approximation of an unknown set specified only by theory T in the language $\mathcal{L}^{ie}_{\mathcal{A}}$ (recall that Theorem 4.2 guarantees the existence of such best approximation).

We will focus on a special class of formulas in $\mathcal{L}^{ie}_{\mathcal{A}}$. A *rule* is every formula φ of the language $\mathcal{L}^{ie}_{\mathcal{A}}$ such that φ is of the form $B \Rightarrow h$, where $B \in \mathcal{L}^{ie}_{\mathcal{A}}$ and h is an atomic formula from $\mathcal{L}^{ie}_{\mathcal{A}}$ (that is, a formula $\mathbf{in}(t)$ or $\mathbf{ex}(s)$ for some $t, s \in \mathcal{L}_I$). We refer to B as the *body* and to h as the *head* of a rule φ. Atomic formulas are special cases of rules (with empty body, which can be interpreted as true formula) as are formulas $\beta \Rightarrow \alpha$, where α and β are atomic formulas in $\mathcal{L}^{ie}_{\mathcal{A}}$.

A rule $B \Rightarrow \mathbf{in}(t)$ captures the following constraint: if a set X satisfies B then it must contain all elements that have property t. A rule $B \Rightarrow \mathbf{ex}(s)$ has a similar interpretation. Thus, in particular, a rule $\mathbf{ex}(s) \Rightarrow \mathbf{in}(t)$ captures the constraint that if a set X does not contain any record from query s then it must contain all records from query t.

In what follows we will consider the class of *rule theories*, that is, theories consisting of rules. We start with rule theories that consist of atomic formulas only. Let $I = \langle \mathcal{U}, \mathcal{A} \rangle$ be an information system. Let T be a set of atomic formulas from $\mathcal{L}^{ie}_{\mathcal{A}}$. Define:

$$L_T = \bigcup \{|t|_I : \mathbf{in}(t) \in T\}, \quad U_T = \mathcal{U} \setminus \bigcup \{|s|_I : \mathbf{ex}(s) \in T\}.$$

Clearly, under any of the finiteness conditions, both L_T and U_T are definable. We have the following straightforward result.

Proposition 5.2. *Let T be a rule theory consisting of atomic formulas of $\mathcal{L}^{ie}_{\mathcal{A}}$. Then, T is consistent if and only if $\langle L_T, U_T \rangle$ is a rough set. Moreover, if T is consistent then $\langle L_T, U_T \rangle$ is the $\preceq_{kn}$-least 3-valued model of T and it coincides with the rough set $\langle \underline{T}, \overline{T} \rangle$.*

Proof: First, assume that T is consistent. Then there is a set X satisfying T. It is easy to see that X satisfies T if and only if $L_T \subseteq X \subseteq U_T$. Thus $L_T \subseteq U_T$, and $\langle L_T, U_T \rangle$ is a rough set. Moreover, clearly, $\langle L_T, U_T \rangle = \langle \underline{T}, \overline{T} \rangle$.

Conversely, if $L_T \subseteq U_T$, then for any t such that $\mathbf{in}(t) \in T$ and for any s such that $\mathbf{ex}(s) \in T$, $|t|_I \cap |s|_I = \emptyset$. But then every set X such that $L_T \subseteq X \subseteq U_T$ is a model of T. It follows that T is consistent. $\square$

We will now extend this result to all rule theories. To this end, we introduce, for each rule theory T, an operator O_T on the lattice $\mathcal{D}_I \times \mathcal{D}_I$ of pairs of definable sets of an information system I (notice that $\mathcal{D}_I \times \mathcal{D}_I$ in addition to rough sets contains additional, "inconsistent" pairs, too). Let T be a rule theory and let R be a a pair of definable sets. Define

$$K(R) = \{\alpha : B \Rightarrow \alpha \in T \quad \text{and} \quad R \models_{I,3} B\}$$

Clearly, $K(R)$ is a rule theory consisting of atomic formulas only. Define

$$O_T(R) = \langle L_{K(R)}, U_{K(R)} \rangle.$$

It is easy to see that under any of the finiteness conditions, for every pair of definable sets R, $O_T(R)$ is also a pair of definable sets, although not always a consistent one.

The fundamental property of the operator O_T is its monotonicity with respect to the ordering $\preceq_{kn}$.

Proposition 5.3. *Let T be a rule theory. Then, the operator O_T is $\preceq_{kn}$-monotone.*

Proof: Let $R_1 \preceq_{kn} R_2$ be two rough sets. We claim that $K(R_1) \subseteq K(R_2)$. Indeed, let $\alpha \in K(R_1)$. Then, there is a rule in T, say $B \Rightarrow \alpha$, such that $R_1 \models_{I,3} B$. Then, by Theorem 5.2, $R_2 \models_{I,3} B$. Consequently, $\alpha \in K(R_2)$.

Next, observe that if T_1, T_2 are two sets of atomic formulas such that $T_1 \subseteq T_2$, then $L_{T_1} \subseteq L_{T_2}$ and $U_{T_2} \subseteq U_{T_1}$. Applying this remark to $K(R_1)$ and $K(R_2)$, we obtain the result. □

Since $\mathcal{D}_I \times \mathcal{D}_I$ is a complete lattice, Knaster-Tarski Theorem [21] implies the following corollary.

Corollary 5.2. *If T is a rule theory, then the operator O_T possesses a $\preceq_{kn}$-least fixpoint.*

The operator O_T has the following intuition. It updates an approximation R by replacing it with the approximation $\langle L_{K(R)}, U_{K(R)} \rangle$. If we iterate O_T starting with $\langle \emptyset, \mathcal{U} \rangle$, in each step (until we reach the fixpoint) we obtain a better approximation to a set X specified by T. We will denote the $\preceq_{kn}$-least fixpoint of O_T by $\langle l_T, u_T \rangle$. Our next result shows that $\langle l_T, u_T \rangle$ approximates the rough set $\langle \underline{T}, \overline{T} \rangle$.

Theorem 5.3. *Let I be an information system satisfying one of the finiteness conditions and let T be a consistent rule theory. Then $\langle l_T, u_T \rangle \preceq_{kn} \langle \underline{T}, \overline{T} \rangle$.*

Proof: Recall that $\langle l_T, u_T \rangle$, the least fixpoint of the operator O_T, is obtained by iterating the operator O_T starting at the least element of $\mathcal{R}_I$, $\langle \emptyset, \mathcal{U} \rangle$. Since I satisfies one of the finiteness conditions, $\langle l_T, u_T \rangle = O_T^n(\langle \emptyset, \mathcal{U} \rangle)$ for some natural number n. By induction on m, we show that for every model X of T, $O_T^m(\langle \emptyset, \mathcal{U} \rangle) \preceq_{kn} \langle X, X \rangle$. This is certainly true for $m = 0$. Assume now that $R = O_T^m(\langle \emptyset, \mathcal{U} \rangle)$ has the property $R \preceq_{kn} \langle X, X \rangle$. We will show that $O_T(R) \preceq_{kn} \langle X, X \rangle$. Consider the formula $\mathbf{in}(t)$ belonging to the set $K(R)$. Then, there is a rule $B \Rightarrow \mathbf{in}(t)$ such that $R \models_{I,3} B$. Consequently, $\langle X, X \rangle \models_{I,3} B$. By Proposition 5.1, $X \models_I B$. Since $B \Rightarrow \mathbf{in}(t)$ belongs to T and since X is a model of T, $X \models_I \mathbf{in}(t)$. Thus, $|t|_I \subseteq X$. We have just proved that whenever $\mathbf{in}(t)$ belongs to $K(R)$, $|t|_I \subseteq X$. It follows that $L_{K(R)} \subseteq X$ and, consequently, that

$$L_{K(R)} \subseteq \bigcap\{X : X \models_I T\} = \underline{T}.$$

Similarly we show that

$$\overline{T} = \bigcup\{X : X \models_I T\} \subseteq U_{K(R)}.$$

Therefore $O_T(R) \preceq_{kn} \langle \underline{T}, \overline{T}\rangle$ and, consequently, $\langle l_T, u_T\rangle \preceq_{kn} \langle \underline{T}, \overline{T}\rangle$. □

Thus, the operator O_T allows us to construct a lower estimate to the best approximation of an unknown set specified by a rule theory. In general, this lower estimate $\langle l_T, u_T\rangle$ is different from the best approximation $\langle \underline{T}, \overline{T}\rangle$. In some cases, however, they coincide.

We say that a formula φ is *positive* if it is built out of atomic formulas by means of conjunctions and alternatives. Thus negation, implication and equivalence symbols are not allowed in positive formulas.

Theorem 5.4. *Let I be an information system satisfying one of the finiteness conditions and let T be a consistent theory whose all rules have positive bodies. Assume that $\langle l_T, u_T\rangle$ is a concrete rough set. Then $\langle l_T, u_T\rangle = \langle \underline{T}, \overline{T}\rangle$.*

Proof: By Theorem 5.3, $\langle l_T, u_T\rangle \preceq_{kn} \langle \underline{T}, \overline{T}\rangle$. Thus, it suffices to show that $\langle \underline{T}, \overline{T}\rangle \preceq_{kn} \langle l_T, u_T\rangle$. First, observe that for every positive formula φ, $X \models_I \varphi$ if and only if $\langle \underline{X}, \overline{X}\rangle \models_{I,3} \varphi$ (an easy proof by induction on the length of φ is omitted).

Let X be a set dense in $\langle l_T, u_T\rangle$, that is, $\underline{X} = l_T$, and $\overline{X} = u_T$ (such a set exists as $\langle l_T, u_T\rangle$ is concrete). Let $\varphi \Rightarrow \alpha$ be a rule in T. Assume that $X \models_I \varphi$. Then, since φ is positive, our observation implies that $\langle \underline{X}, \overline{X}\rangle \models_{I,3} \varphi$. Since $\langle l_T, u_T\rangle = \langle \underline{X}, \overline{X}\rangle$, $\langle l_T, u_T\rangle \models_{I,3} \varphi$. Recall that the rough set $\langle l_T, u_T\rangle$ is the fixpoint of the operator O_T. Thus, $\langle l_T, u_T\rangle \models_{I,3} \alpha$. By Theorem 5.1, it follows that $X \models_I \alpha$ and, consequently, X is a model of T. Hence, $\langle \underline{T}, \overline{T}\rangle \preceq_{kn} \langle X, X\rangle$. Thus, by Theorem 3.3, we obtain $\langle \underline{T}, \overline{T}\rangle \preceq_{kn} \langle l_T, u_T\rangle$. □

As noticed above, a rule theory does not need to be consistent. In fact, even a theory consisting of atoms need not to be consistent. It should be clear that checking if a theory T consisting of atoms is consistent can be done by a number of calls to satisfiability engine that is proportional to the square of the size of T.

If a theory T consists of rules with positive body, then, by computing the fixpoint of the operator O_T we arrive at a pair of definable sets. If that pair is not consistent, T itself is not consistent. If that pair is consistent, and if the resulting rough sets is concrete, then we computed the rough set $\langle \underline{T}, \overline{T}\rangle$. It is quite clear that this computation requires only a polynomial number of calls to the satisfiability engine.

There are classes of theories that are guaranteed to be consistent. One example of such theories is the class of *safe* rule theories.

A theory T consisting of rules is *safe* over I if for every formula $\mathbf{in}(t)$ occurring as the head of a rule in T and every formula $\mathbf{ex}(s)$ occurring as a head of a rule in T, $|t \cdot s|_I = \emptyset$.

Corollary 5.3. *If I is an information system then any positive safe rule theory T over I is consistent. Thus, if $\langle l_T, u_T\rangle$ is concrete, then it is equal to $\langle \underline{T}, \overline{T}\rangle$.*

Notice, however, that checking safeness is expensive. It requires quadratic (in the cardinality of T) number of calls to the satisfiability engine. Thus, given a rule theory, rather than to check its safeness it is, in general, better to compute the fixed point first, and then check its consistency at the very end.

6. Problems and future directions

The approach to rough sets proposed in this paper opens several interesting research directions. First, let us note that Pawlak's rough sets or rough sets as defined in this paper may have very complex descriptions. That is, the terms of the language $\mathcal{L}_{\mathcal{A}}$ defining them may have exponential length with respect to the number of atomic terms they involve. Thus, we should not only be interested in finding approximations to unknown sets but also in finding *short* approximations.

To formalize the concept of a "short" description we will now introduce the notion of k-definability. Let $I = \langle \mathcal{U}, \mathcal{A}\rangle$ be an information system. A definable set X is *k-definable* if there is a term $t \in \mathcal{L}_{\mathcal{A}}$ of length at most k and such that $|t|_I = X$. A rough set $\langle L, U\rangle$ is k-definable if both L and U are k-definable.

Asking simply for a short approximation does not lead to interesting research problems. After all the trivial approximation $\langle \emptyset, \mathcal{U}\rangle$ approximates all sets and has a very short description. Interesting problems arise when the requirement for a short description is combined with a requirement for a high precision of the approximation. Pawlak [15] studied several precision measures. For instance, the *tightness* of an approximation $\langle L, U\rangle$ can be measured by the ratio

$$\frac{size(U \setminus L)}{size(U)}.$$

We can now formulate the following basic problem on the trade-off between length of an approximation and its tightness. Given integers k, l and m, and given a theory T in the language of inclusion-exclusion, is there a rough set R such that R approximates all sets satisfying T, R is k-definable and the tightness of R is at most l/m. Both theoretical and algorithmic results on this problem are of significant practical importance.

Another interesting research direction with many promising applications in the area of data mining is related to an observation that the language of inclusion-exclusion is only the first step towards the language for specifying unknown sets. In the language of inclusion-exclusion unknown sets are described in terms of definable sets which they contain or which they are disjoint with. However, as demonstrated by Proposition 4.2, the language of inclusion-exclusion does not allow us to talk about the sizes of definable sets. In particular, in the language of inclusion-exclusion we cannot formulate requirements that an unknown set intersects with a given definable set on at least (at most) k elements. It is important to generalize the language of inclusion-exclusion to allow one to formulate also numeric constraints on the unknown sets.

Applications in data mining, in particular OLAP applications, may require such an extension of the language [4]. Once an appropriate generalization is proposed, a theory similar to that presented in the present paper should be developed.

Acknowledgements

The authors acknowledge helpful conversations with M. Denecker, W.W. Koczkodaj, Z. Pawlak, and A. Skowron. This work was partially supported by the NSF grants CDA-9502645 and IRI-9619233 as well as the US ARO contract DAAH 04-96-1-0398.

References

[1] Abiteboul, S., Hull, R., and Vianu, V., *Foundations of Databases.* Addison Wesley, 1995.

[2] Buszkowski, W. Approximation spaces and definability in incomplete information systems, in: [16], pp. 115–122.

[3] Chlebus, R.S., Nguyen, S.H., On Finding Optimal discretizations for Two Attributes, in [16], pp. 537–544.

[4] Chaudhuri, S., Dayal, U., An overview of data warehousing and OLAP technology, *SIGMOD Record* 26 (1997) pp. 65-74.

[5] Fitting, M.C., Fixpoint semantics for logic programming – a survey. *Theoretical Computer Science*, 1999. To appear.

[6] Ginsberg, M.L., Multivalued Logics: a uniform approach to reasoning in artificial intelligence, *Computational Intelligence*, 4 (1988), pp. 265-316.

[7] Koczkodaj, W.W., Marek V.W. and Orlowski, M. Myths about Rough Set Theory. *Communications of the ACM* 41 (1998), pp. 102–103.

[8] Kleene, S.C., *Introduction to Metamathematics*, Van Nostrand, 1952.

[9] Kuratowski, K., Mostowski, A., *Set Theory*, North Holland, 1982.

[10] Marek, W., Pawlak, Z., Rough Sets and Information Systems, *Fundamenta Informaticae* 7 (1984), pp. 105-115.

[11] Marek, W., Pawlak, Z., Information storage and retrieval systems, mathematical foundations, *Theoretical Computer Science* 1 (1976), pp. 331–354.

[12] Mitchell, T.M., *Version Spaces: An Approach to Concept Learning*, Ph.D. Dissertation, Stanford University, 1977.

[13] Orlowska, E. (ed.), *Incomplete Information: Rough Set Analysis*, Physica Verlag, 1997.

[14] Pawlak, Z. Rough Sets, *International Journal of Computer and Information Sciences* 11 (1982), pp. 341-356.

[15] Pawlak, Z. *Rough Sets, Theoretical Aspects of Reasoning about Data*, Kluwer, 1991.

[16] Polkowski, L., and Skowron, A. (eds.), *Rough Sets and Current Trends in Computing*, Springer Lecture Notes in Artificial Intelligence 1424, Springer-Verlag, 1998.

[17] Pal, S.K., and Skowron, A. (eds.), *Rough Fuzzy Hybridization - A New Trend in Decision-Making*, Springer-Verlag, 1999.

[18] Rauszer, C., Skowron, A., The discernibility matrices and functions in information systems, In: [20] pp. 311–362.

[19] Sikorski, R. *Boolean Algebras*, Springer-Verlag, 1963.

[20] Słowiński, R. (ed.) *Intelligent Decision Support: Handbook of Applications and Advances in Rough Set Theory*, Kluwer, 1992.

[21] Tarski, A. A lattice theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics* 5 (1955), pp. 285–309.

[22] Ullman, J.D. *Principles of Database and Knowledge Base Systems*, I/II, Computer Science Press, 1988 and 1989.

[23] Ziarko, W.P. (ed.), *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer-Verlag, 1988.